УКРАЇНСЬКИЙ АНТАРКТИЧНИЙ ЖУРНАЛ
УАЖ, №15, 213—220 (2016)

УДК 004.62

КЛАСИФІКАЦІЯ ДАНИХ ДЕРЖАВНОЇ УСТАНОВИ НАЦІОНАЛЬНИЙ АНТАРКТИЧНИЙ НАУКОВИЙ ЦЕНТР

Л. С. Глоба[1], Р. Л. Новогрудська[1], А. А. Сидоренко[1], А. Березкіна[2]

[1] *Національний технічний університет України «Київський політехнічний інститут», Інститут телекомунікаційних систем, пр. Індустріальний, 2, Київ, Україна 03056, e-mail: lgloba@its.kpi.ua, rinan@ukr.net, liddell526@gmail.com*
[2] *Державна установа Національний антарктичний науковий центр МОН України*

Реферат. В роботі запропоновано підхід до створення сховища антарктичних даних. Сховище антарктичних даних є середовищем, що дозволяє зберігати та отримувати доступ до усіх первинних («сирих» даних) та вторинних (результатів досліджень) даних. Таке сховище становить ядро Національного центру антарктичних даних. Досліджено проблему класифікації великого об'єму даних, котрі повинні розміщуватися на порталі Державної установи Національний антарктичний науковий центр (ДУ НАНЦ). Розглянуто основні підходи до систематизації інформації.
В результаті аналізу даних, що містяться у вже існуючих на даний момент системах зберігання інформації ДУ НАНЦ, та даної предметної області в цілому запропоновано системну класифікацію даних ДУ НАНЦ, що повністю описує весь масив представленої інформації. Запропоновано макет Національного центру антарктичних даних, описано інтерфейс доступу до сховища антарктичних даних та структуру такого сховища.

Ключові слова: сховище антарктичних даних, класифікація, портал ДУ НАНЦ, результати антарктичних досліджень, Національний центр антарктичних даних.

Классификация данных Государственного учреждения Национальный антарктический научный центр
Л. С. Глоба, Р. Л. Новогрудская, А. А. Сидоренко, А. Березкина

Реферат. В работе предложен подход к созданию хранилища антарктических данных. Хранилище антарктических данных является средой, позволяющей хранить и получать доступ ко всем первичным («сырым» данным) и вторичным (результатам исследований) данным. Такое хранилище составляет ядро Национального центра антарктических данных. Исследована проблема классификации большого объема данных, которые должны размещаться на портале Государственного учреждения Национальный антарктический научный центр (ГУ НАНЦ). Рассмотрены основные подходы к систематизации информации.
В результате анализа данных, которые хранятся в уже существующих на данный момент системах хранения информации ГУ НАНЦ, и данной предметной области в целом, предложено системную классификацию данных ГУ НАНЦ, полностью описывающую весь массив представленной информации. Предложено макет Национального центра антарктических данных, описано интерфейс доступа к хранилищу антарктических данных и структуру такого хранилища.

Ключевые слова: хранилище антарктических данных, классификация, портал ГУ НАНЦ, результаты антарктических исследований, Национальный центр антарктических данных.

Data classification State institution National Antarctic Scientific Center
L. S. Globa, R. L. Novogrudska, A. A. Sidorenko, A. Berezkina

Abstract. The paper presents an approach to creating a repository of Antarctic data. Antarctic data storage is a medium that allows you to store and access all primary («raw» data) and secondary (research results) data. This repository is the core of the National Antarctic Data Center. The problem of classifying large amounts of data that must be displayed on the portal of the State institution National antarctic scientific center (SI NASC). The basic approach to the systematization of information.
An analysis of the data contained within existing systems currently storage of SI NASC, and the domain of the proposed system in general classification SI NASC data that fully describes the entire array of provided information. A model of the National antarctic data center, describes the interface for accessing antarctic data storage and structure of the repository.

Keywords: Antarctic data storage, classification, portal of SI NASC, the results of Antarctic research, National Antarctic Data Center.

1. Вступ

Протягом останніх двадцяти років Україна є однією з більш ніж двадцяти країн, які проводить дослідження в Антарктиді. За цей час дослідження охопили понад десять наукових напрямів, основними серед яких є: вивчення магнітного поля Землі, дослідження іоносфери, метеорологічні дослідження, велика кількість медичних досліджень, також ведеться активна робота по дослідженню флори і фауни Антарктиди (Новогрудська Р. Л., 2015, Глоба Л. С., 2011). У результаті проведення досліджень накопичено великий масив інформації, котрий натепер доступний лише вузькому колу науковців. Помітним негативним чинником процесу доступу до знань є недостатня систематизація і слабка структурованість великих обсягів інформації в існуючих на даний момент системах зберігання. Отже, однією з задач при розробці Національного центру антарктичних даних стало проведення класифікації антарктичних даних відповідно до напрямків досліджень, що проводяться на станції «Академік Вернадський».

Єдине інформаційне середовище даних антарктичних досліджень складається з трьох основних компонент:

- портал ДУ НАНЦ – середовище для збереження та представлення інформації, що стосується роботи ДУ НАНЦ (інформація про напрямки досліджень, експедиції на станцію Академік Вернадський, видавничу діяльність, нормативні документи, архів новин та ін.);
- модуль опису результатів антарктичних досліджень – середовище, що дозволяє не лише зберігати кінцеві результати антарктичних досліджень, а й задавати та зберігати специфічні метаописи кожного результату дослідження. Таке середовище дає змогу представити результати антарктичних досліджень згідно з вимогами SCAR. Використання такого модулю допомагає передавати результати антарктичних досліджень на портал NASA;

Рис. 1. Взаємодія компонентів єдиного інформаційного середовища даних антарктичних досліджень

- сховище антарктичних даних – середовище, що дозволяє зберігати та отримувати доступ до усіх первинних («сирих» даних) та вторинних (результатів досліджень) даних. Таке сховище становить ядро Національного центру антарктичних даних.

Національний центр антарктичних даних є інтеграцією двох компонент: компоненти опису результатів антарктичних досліджень для передачі на портал NASA та сховище антарктичних даних (рисунок 1).

Розробка та впровадження Національного центру антарктичних даних вирішить завдання збору, обробки, передачі та представлення даних результатів антарктичних досліджень.

Основним завданням, вирішення якого описано у пропонованій роботі, є створення сховища антарктичних даних, що забезпечить зберігання усієї інформації за всіма напрямками досліджень, що проводяться в ДУ НАНЦ та суміжних організаціях.

Головними завданнями, які необхідно вирішити в рамках створення сховища антарктичних даних, є:

1) запропонувати класифікацію антарктичних даних;

2) розробити макет Національного центру антарктичного даних (сховище антарктичних даних, інтегроване зі сховищем результатів антарктичних досліджень для порталу NASA);

3) розробити структуру сховища даних за напрямками досліджень;

4) встановити автоматизований зв'язок між порталом ДУ НАНЦ та сховищем;

5) розробити технологію підтримки гетерогенної інфраструктури (порталу ДУ НАНЦ та сховища) для зберігання даних.

У даній статті описана реалізація перших трьох завдань.

2. Підхід до класифікації великих об'ємів даних

У ході виконання роботи постала задача класифікації великих об'ємів даних, які мали різну форму (документи, зображення, медіа файли, файли з специфічним розширенням).

Важливою рисою ефективної роботи є швидкий доступ до необхідних ресурсів. Однією з задач, що постала перед нами, стала систематизація матеріалів ДУ НАНЦ.

Систематизована інформація значно краще сприймається і запам'ятовується людиною. Знайти потрібні відомості в систематизованому сховищі набагато простіше й швидше, обробляти систематизовані дані легше. Залежно від поставленої задачі систематизація інформації може зводитися до її аналізу, структурування, упорядкування, формалізації, класифікації, кластеризації, типологізації або до комбінації декількох з цих процедур.

Загалом, під систематизацією інформації розуміється свого роду класифікація всіх документів за різними групами. Систематизація інформації включає:

— методи пошуку і накопичення інформації;

— класифікацію та індексування інформації;

— способи доступу до інформації;

— способи подання інформації.

Необхідно обрати найбільш зручний для даної предметної області метод систематизації інформації, той чи інший тип класифікації (або сукупність таких типів). Найчастіше всі дані розподіляються відповідно до номінальної, предметної, тематичної, хронологічної, авторської і архівної класифікації:

1) номінальна систематизація – розподіл документів по їх типу (текстові звіти, бінарні файли, фото- та відео-матеріали, Matlab-графіки та інші файли);

2) предметна систематизація – розподіл за належністю документа до якоїсь конкретної справи (наприклад, матеріали, отримані під час експедиції: фото- та відео-файли, звіти та інші файли);

3) тематична систематизація – групування даних за загальною тематикою (біологічні, медичні дослідження, дослідження іоносфери та інші);

4) хронологічна систематизація інформації – розподіл документів за датою їх створення;

5) авторська систематизація – за прізвищем вченого чи фахівця, що є автором документа;

6) архівна − за термінами зберігання документації.

Систематизація інформації передбачає обробку інформації з метою приведення її до певного виду та представлення інформації, що дозволяє користувачеві у певний спосіб використовувати інформацію. В результаті обробки інформація розміщується в певному порядку, набуває якоїсь завершеної форми, наповнюючись певним змістом і значенням. Обробка інформації створює образи, форми, які користувач може розпізнати і які розуміє певним чином. При цьому відбувається процес

зведення комплексу даних до спрощених структурованих категорій, кожна з яких займає певне місце в загальному масиві даних (Методы, 2016).

Варто відмітити один з важливих аспектів поставленої перед нами задачі: при розробці автоматизованих інформаційних систем (починаючи від інформаційно-пошукових систем, закінчуючи системами робототехніки) процеси сприйняття, кодування, передачі і зберігання інформації мають певну специфіку. Враховуючи, що інформація, розміщена на ресурсі, специфічна і представлена файлами різного формату, необхідно було підібрати оптимальний варіант структурування даних та спосіб їх збереження на порталі.

Існують різні способи структурування даних:

— модель даних на основі записів;

— об'єктно-орієнтована модель;

— фізична модель.

Кожен метод має свої особливості розбиття даних.

Фізичні моделі даних описують те, як дані зберігаються на комп'ютері, представляючи інформацію про структуру записів, їх упорядкованість та існуючі шляхи доступу (Базы, 2000).

Об'єктно-орієнтована модель даних основана на поняттях об'єктно-орієнтованого програмування: дані представлені у вигляді об'єктів, для яких визначені не лише властивості, а й методи. Дана модель використовується для зберігання даних, що мають складну структуру та дозволяють визначити функції їх обробки. Проте натепер об'єкто-орієнтовані бази даних лише починають розвиватись, вони досить складні для розуміння користувача та в більшості випадків мають занадто низьку швидкодію.

При використанні моделі даних на основі записів база даних складається з декількох записів фіксованого формату, які можуть мати різні типи. Кожен тип запису визначає фіксовану кількість полів, кожне з яких має фіксовану довжину.

Існують три основні типи логічних моделей даних на основі записів:

— реляційна модель даних;

— мережева модель даних;

— ієрархічна модель даних.

Реляційна модель характерна тим, що дані представлені у вигляді набору логічно зв'язаних таблиць – відношень. Ця модель сьогодні є основною формою зберігання даних. Реляційна модель має ряд переваг, зокрема: простота і доступність для розуміння користувачем, повна незалежність структури даних від прикладних програм та ін. Проте далеко не всі дані можуть бути представлені у вигляді таблиць: оскільки дані ДУ НАНЦ мають різний формат і подекуди специфічне розширення, ми були змушені відмовитись від використання реляційної моделі.

Обрана для організації даних ДУ НАНЦ ієрархічна модель даних (рис. 2) будується за принципом підпорядкованості між елементами даних і характеризується деревоподібною структурою, яка складається з вузлів (сегментів) і дуг (гілок). Дерево в ієрархічній структурі впорядковане за правилами його сегментів і гілок: на верхньому рівні — один кореневий (вихідний) сегмент; сегмент другого рівня, породжений, залежить від першого, вихідного; доступ до кожного породженого (крім кореневого) здійснюється через його вихідний сегмент; кожний сегмент може мати кілька примірників конкретних значень елементів даних, а кожний елемент породженого сегмента пов'язаний із примірником вихідного і створює один логічний запис; примірник породженого сегмента не може існувати самостійно, тобто без кореневого сегмента; при вилученні примірника кореневого сегмента вилучаються також усі підпорядковані та взаємопов'язані з ним примірники породжених сегментів (Глоба Л. С., 2007).

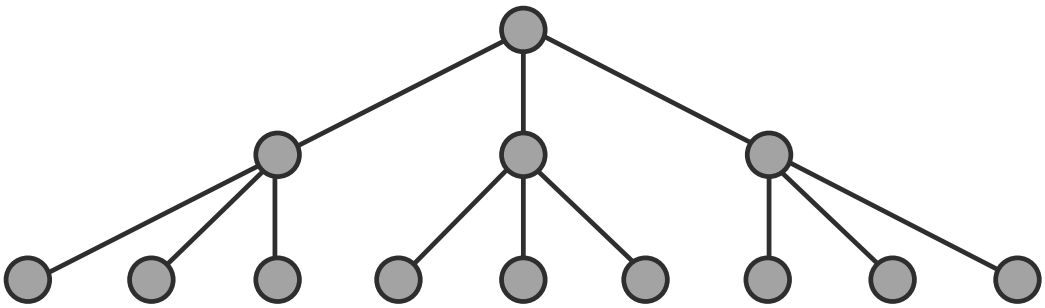

Рис. 2 Ієрархічна модель організації даних

3. Класифікація антарктичних даних за напрямками досліджень

У процесі розробки сховища антарктичних даних виникла задача розробки класифікації антарктичних даних за наступними напрямками досліджень:

1) геомагнетизм;
2) метеорологічні дослідження;
3) дослідження іоносфери;
4) біологічні дослідження;
5) медичні дослідження.

Головним інструментом при розробці класифікації даних, без сумніву, є спілкування з фахівцями, які працюють у даній організації, адже вони, як ніхто інший, розуміють, яка саме інформація необхідна для їхньої роботи, та яку роль відіграють ті чи інші дані на тлі загальної інформаційної картини. Тому, щоб зрозуміти, як саме організувати дані, котрі повинні зберігатися на порталі ДУ НАНЦ, у процесі роботи проводилися зустрічі з представниками ДУ НАНЦ, що працюють за кожним з зазначених вище напрямків. Для кожного напрямку було детально оговорено вимоги до класифікації даних та проаналізовано дані, розміщені в існуючій на даний момент системі зберігання даних ДУ НАНЦ.

Варто зазначити, що початкова задача передбачала створення сховища первинних даних, отриманих безпосередньо зі станції. В результаті роботи стало зрозумілим, що необхідне створення загального сховища інформаційних надбань ДУ НАНЦ, що включає як і так звані «сирі» дані, так і річні звіти та публікації науковців та фахівців ДУ НАНЦ.

Як результат, згідно з представленнями користувачів, для кожного з напрямків досліджень було розроблено унікальні концептуальні моделі даних, які максимально відповідають поставленим вимогам.

Класифікація антарктичних даних є основою для структуризації та систематизації даних антарктичних досліджень. Саме згідно з розробленою класифікацією дані відображаються для кінцевого користувача при роботі з інтерфейсом відповідного модулю на порталі (інтерфейс, що дозволяє здійснювати пошук, перегляд та збереження даних антарктичних досліджень) (Структурирование, 2016). Класифікація антарктичних даних виконувалася на основі рекомендацій науковців, які є відповідальними за певний напрям наукових досліджень.

В рамках кожного напрямку виділено підтипи досліджень (від трьох до п’яти рівнів углиб ієрархії).

Наприклад, так представлено класифікацію даних за напрямками досліджень Геомагнетизм та Біологія:

Геомагнетизм

- *Оригінальні дані*
  - ⇒ Роки
    - ✓ Дані варіометра
      - ➢ LEMI-08 №2 (MDZ)
      - ➢ LEMI-08 №16 (XOY)
    - ✓ POS – 1
    - ✓ Абсолютні вимірювання
- *Оброблені дані*
  - ⇒ Роки

Біологічні дослідження

- ✓ Відео
  - ➢ 1 УАЕ (1996)
  - ➢ 2 УАЕ(1997)
  - ➢ ……
- ✓ Звіти
  - ➢ Звіти УАЕ з станції
    - ○ Річні звіти → Роки

- Місячні звіти → Роки → Місяці
- ➢ Звіти сезонних експедицій → Роки
- ➢ Звіти виконувачів НТР → Роки
  - Харківський національний університет ім.Каразіна
  - Інститут молекулярної біології і генетики НАН України
  - УНЦ «Інститут біології» КНУ ім.Шевченка

✓ Статті
- ➢ Тварини
  - Безхребетні тварини
  - Хребетні тварини
- ➢ Рослини
- ➢ Екологія
- ➢ Віруси
- ➢ Інше

✓ Фотографії
- ➢ 1 УАЕ (1996)
- ➢ 2 УАЕ(1997)

4. Організація даних на порталі ДУ НАНЦ

Наступним етапом роботи стала технічна реалізація розробленої класифікації даних та розміщення її на порталі ДУ НАНЦ. Така реалізація задає інтерфейс для відображення структури сховища антарктичних даних за напрямками досліджень.

На порталі ДУ НАНЦ відповідно до затвердженої класифікації даних було розроблено зручну ієрархію даних, яка згодом за необхідності може бути з легкістю відредагована користувачами, що мають відповідні права.

Доступ на портал ДУ НАНЦ мають лише зареєстровані користувачі. Для дослідників, що працюють за кожним з напрямків, було створено облікові записи, котрі занесено до Матриці доступу співробітників. Матриця доступу описує, в якій галузі працює даний користувач. Відповідно, він має право на додавання, видалення та редагування даних лише за своїм напрямком досліджень.

Класифікація даних представлена у вигляді деревовидного списку. Верхній рівень класифікації містить п’ять категорій відповідно до напрямків досліджень, що проводяться фахівцями ДУ НАНЦ (рис. 3).

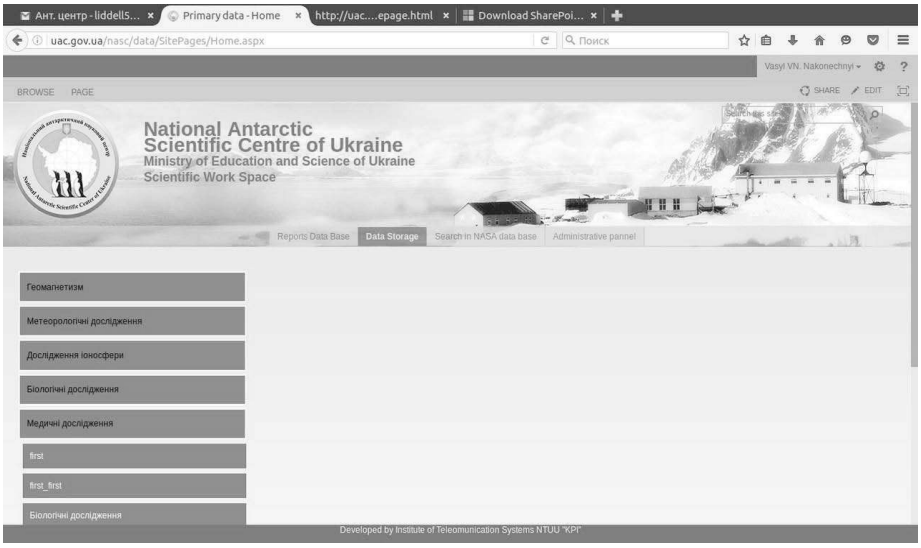

Рис. 3. Класифікація даних на порталі ДУ НАНЦ: верхній рівень організації даних

Як приклад, розглянемо класифікацію даних за напрямком досліджень Геомагнетизм. Як зазначалось раніше, дані було доцільно класифікувати на первинні дані, отримані безпосередньо зі станції, та вже опрацьовані. Класифікація описує другий рівень ієрархії даних.

Оскільки апаратне забезпечення, що використовується на станції для проведення відповідних вимірів, за двадцять років роботи станції змінювалося, було прийнято рішення провести подальшу класифікацію первинних даних по роках і для кожного року вказати, які саме виміри проводились і які технології при цьому використовувались (рис. 4).

Первинні дані у сфері геомагнетизму − це бінарні файли, файли, та файли зі специфічним розширенням, котрі вимагають для перегляду певного програмного забезпечення.

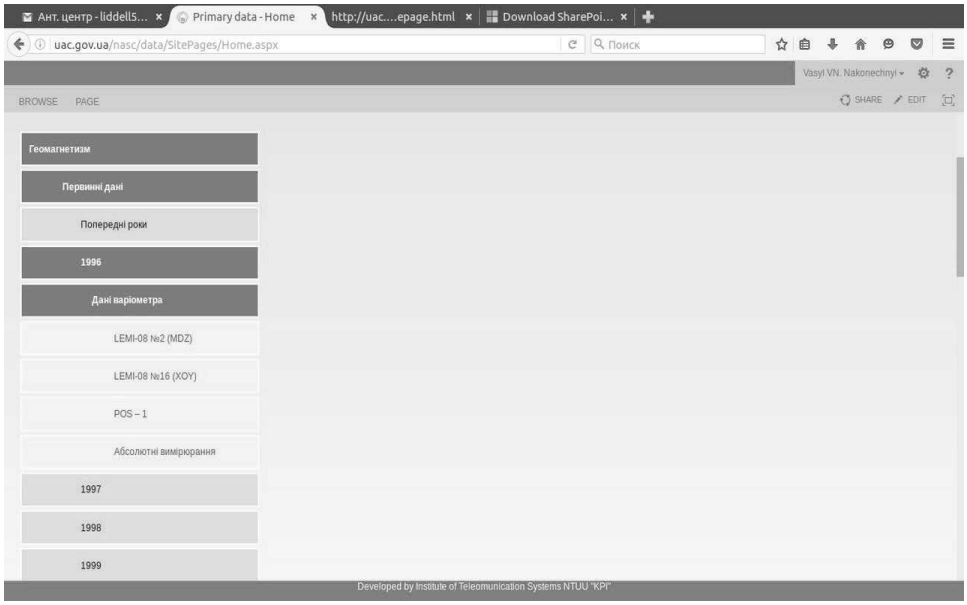

Рис. 4. Приклад класифікації первинних даних на порталі ДУ НАНЦ

Оброблені дані також класифікуються по роках: це, загалом, річні та місячні звіти, звіти експедицій (рис. 5).

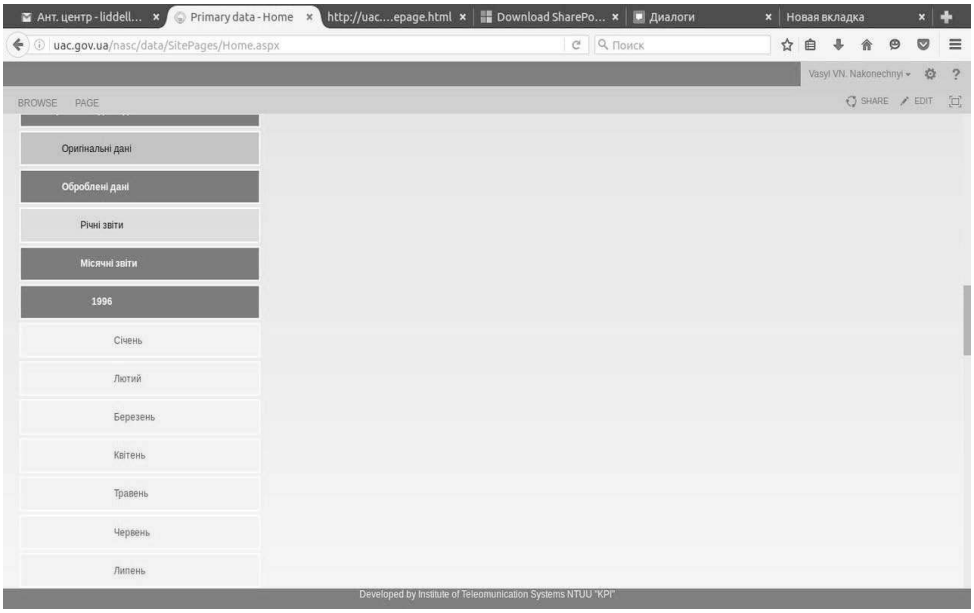

Рис. 5. Приклад класифікації оброблених даних на порталі ДУ НАНЦ

5. Висновок

У роботі запропонована класифікація антарктичних даних, які будуть розміщуватись на порталі ДУ НАНЦ з метою підвищення ефективності доступу до результатів досліджень, проведених на станції «Академік Вернадський».

Проведено систематизацію даних антарктичних досліджень. Відповідно до запропонованої класифікації розроблено структуру сховища антарктичних даних, що забезпечить зберігання усієї інформації за всіма напрямками досліджень, які проводяться в ДУ НАНЦ та суміжних організаціях.

На порталі ДУ НАНЦ розроблено інтерфейс користувача та відповідне функціональне меню, за допомогою якого науковці зможуть з легкістю знайти необхідні дані за різними напрямками досліджень.

Список літератури

1. **Базы** данных: Проектирование, реализация и сопровождение. Теория и практика // М. : Вильямс, 2000. − 1093 с. Первое издание. Учебник.
2. **Глоба Л. С.** Створення та обробка баз даних / Л. С. Глоба, М. Ю. Терновой // К., 2007.
3. **Глоба Л. С.** Создание единого информационного пространства данных антарктических исследований / Л. С. Глоба, І. В. Мороз, Р. Л. Новогрудская, К. С. Мочалкина, І. О. Кузін // Український Антарктичний Журнал, № 10−11, 2011, С. 343−351.
4. **Методы** систематизации информации [Електронний ресурс]. – Електрон. текстові дані. – Режим доступу: новыйспособ.рф/методы-систематизации-информации. Дата доступу: 14.12.2016
5. **Новогрудська Р. Л.** Системний підхід до моделювання порталу «Національний центр антарктичних даних» / Р. Л. Новогрудська, Н. В. Дерманська // Український Антарктичний Журнал. − №14. – 2015, С. 238—245.
6. **Структурирование** данных: что делать с интернетом? [Електронний ресурс]. – Електрон. текстові дані. – https://habrahabr.ru/post/143577/ Дата доступу: 09.12.2016